UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

KAI BIRD, JONATHAN ALTER, MARY BLY, VICTOR LAVALLE, EUGENE LINDEN, DANIEL OKRENT, HAMPTON SIDES, JIA TOLENTINO, RACHEL VAIL, SIMON WINCHESTER, AND ELOISA JAMES, INC., individually and on behalf of others similarly situated, Plaintiffs, v. MICROSOFT CORPORATION, Defendant.	**ECF CASE** No. 1:25-cv-05282 **CLASS ACTION COMPLAINT** **JURY TRIAL DEMANDED**

Plaintiffs Jonathan Alter, Mary Bly, Victor LaValle, Rachel Vail, Kai Bird, Eugene Linden, Daniel Okrent, Hampton Sides, Jia Tolentino, Simon Winchester, and Eloisa James, Inc., on behalf of themselves and all other similarly situated, for their complaint against Defendant Microsoft Corporation ("Microsoft" or "Defendant"), allege as follows:

NATURE OF THE CASE

1. Apart from its well-publicized partnership with OpenAI, Microsoft has taken independent action to develop its own multi-billion dollar artificial intelligence technology by taking the combined works of humanity without permission. Rather than pay for intellectual property, they knowingly ignored the laws protecting copyright in two respects: downloading pirated works, and training on copyright registered works without consent or compensation.

2. Plaintiffs, copyright holders of a broad array of works, bring this action under the Copyright Act seeking redress for Microsoft's flagrant and harmful infringements of Plaintiffs' registered copyrights.

3. Microsoft is the developer of the "Turing"-line of Large Language Models or LLMs. Microsoft released a private demo of the first Turing LLM, named "Turing-NLG," in February 2020. In October 2021, Microsoft published a blog post, introducing the significantly larger successor model, the "Megatron-Turing Natural Language Generation model (MT-NLG)" (hereinafter "Megatron" or "Megatron LLM"). LLMs are algorithms designed to output human-seeming text responses to users' prompts and queries. To generate text output that resembles human expression, LLMs must be trained on a large, diverse corpus of text written by humans.

4. To train its Megatron LLM, Microsoft copied a notorious collection of approximately 200,000 pirated books known as "Books3"[1] and fed them into its LLM. The Books3 dataset contains Plaintiffs' works and Microsoft copied their works for LLM training without authorization when it copied Books3.

5. Microsoft could have "trained" Megatron on works in the public domain. It could have paid a reasonable licensing fee to use copyrighted works. But either of those would have taken longer and cost more money than the option Microsoft chose: to train its LLM without permission and compensation as if the laws protecting copyrighted works did not exist. Yet the United States Constitution protects the fundamental principle that creators, like authors, deserve compensation for their works, and the Copyright Act grants "a bundle of exclusive rights" to creators, including "the rights to reproduce the copyrighted work[s]." *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 526 (2023).

[1] *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model*, Microsoft Research Blog (Oct. 11, 2021), *published at* https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/ (last accessed Mar. 17, 2025).; Alex Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative AI,* The Atlantic (Aug. 19, 2023), https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/.

6. The end result is a computer model that is not only built on the work of thousands of creators and authors, but also built to generate a wide range of expression that mimics the syntax, voice, and themes of the copyrighted works on which it was trained.[2]

7. Upon information and belief, Microsoft knew that it needed a license to use Plaintiffs' works to train its LLM, and thus knew that its conduct was infringing.

8. Indeed, Microsoft, as of November 2024 entered into a licensing deal for AI training data, namely books, with publisher HarperCollins. This license provides the right to use a given work in training for three years in exchange for a payment of $5,000 split evenly between the author and the publisher.[3] Microsoft thus acknowledges the right of creators to be compensated when their works are used for the training of AI models, including LLMs.[4]

9. Microsoft's intentional decision to use pirated libraries allowed it to gain huge advantages in the timing and efficiency of its LLMs. For example, Microsoft noted in a blog post that "[t]he innovations of DeepSpeed and Megatron-LM will benefit existing and future AI model development and make large AI models cheaper and faster to train. We look forward to how MT-NLG will shape tomorrow's products and motivate the community to push the boundaries of NLP even further."[5] Meanwhile, its use of pirated libraries helped sustain and foster rampant copyright violations by keeping these pirated libraries in business and providing them a seal of approval.

---

[2] *See, e.g.*, Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model, Shaden Smith et al., https://arxiv.org/pdf/2201.11990 (last visited June 22, 2025).

[3] Andrew Albanese and Jim Milliot, *Agents, Authors Question HarperCollins AI Deal*, Publisher's Weekly (Nov. 19, 2024), *available at* https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/96533-agents-authors-question-harpercollins-ai-deal.html (last visited Mar. 2, 2025)

[4] Hannah Miller and Dina Bass, *Microsoft Signs AI Learning Deal With News Corp.'s HarperCollins*, Bloomberg (Nov. 19, 2024), *available at* https://www.bloomberg.com/news/articles/2024-11-19/microsoft-signs-ai-learning-deal-with-news-corp-s-harpercollins (last visited Feb. 7, 2025).

[5] https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/ (last visited June 20, 2025).

10. In training its models, Microsoft reproduced copyrighted texts to exploit precisely what the Copyright Act was designed to protect: the elements of protectible expression within them, like the choice and order of words and sentences, syntax, flow, themes, and paragraph and story structure. In other words, the goal of the training process was to teach the model to learn how words fit together grammatically, how words work together to form higher-level ideas, and how sequences of words form structured thoughts.[6] In other words, by training its models on certain works, Microsoft copied the works' expression so that the models could memorize, mimic, and paraphrase that expression. Defendants copied and data-mined the works of writers, without permission or compensation, to build a machine that is capable (or, as technology advances, will soon be capable) of performing the same type of work for which these writers would be paid.

11. Microsoft's commercial gain has come at the expense of creators and rightsholders like Plaintiffs and members of the Class. A person who reads a book typically buys it from a store. But Microsoft did not even do that. Instead, Microsoft took these works; it made unlicensed copies of them; and it used those unlicensed copies to digest and analyze the copyrighted expression in them, all for commercial gain.

12. Plaintiffs seek to represent a class of copyright holders whose books works were downloaded and/or used to train Microsoft's artificial intelligence models ("Class"). Plaintiffs, on behalf of themselves and the Class, seek damages from Microsoft for its largescale infringement of their copyrighted works, as well as injunctive relief.

JURISDICTION & VENUE

13. The Court has subject matter jurisdiction under 28 U.S.C. §§ 1331 and 1338(a) because this action arises under the Copyright Act of 1976, 17 U.S.C. § 101, *et seq*.

[6] Fred von Lohmann, response to Notice of Inquiry and Request for Comment 5, (Oct. 30, 2023), *available at* https://downloads.regulations.gov/COLC-2023-0006-8906/attachment_1.pdf.

14. Microsoft maintains offices and employs personnel in New York and has purposely availed themselves of the privilege of conducting business in New York.

15. The harms flowing from Microsoft's copyright infringement occurred, in substantial part, in this District. Plaintiffs Bird, Bly, LaValle, Linden, Robinson, Tolentino, Vail, Winchester, and Eloisa James, Inc., are citizens of New York.

16. Venue is proper under 28 U.S.C. § 1400(a) because Microsoft or its agents reside or may be found in this District.

THE PARTIES

A. Plaintiffs

17. Plaintiff Jonathan Alter is an author and a resident of New Jersey.

18. Plaintiff Kai Bird is an author and a resident of New York.

19. Plaintiff Mary Bly is an author and a resident of New York.

20. Plaintiff Victor LaValle is an author and a resident of New York.

21. Plaintiff Eugene Linden is an author and a resident of New York.

22. Plaintiff Daniel Okrent is an author and a resident of Massachusetts.

23. Plaintiff Hampton Sides is an author and a resident of New Mexico.

24. Plaintiff Jia Tolentino is an author and a resident of New York.

25. Plaintiff Rachel Vail is an author and a resident of New York.

26. Plaintiff Simon Winchester is an author and a resident of New York.

27. Eloisa James, Inc. is a loan-out corporation owned by Plaintiff Mary Bly.

28. Eloisa James, Inc. is a New York corporation with its principal place of business in New York.

29. The registration information for the infringed works of Plaintiffs is identified in Exhibit A to this Complaint.

B. Microsoft

30. Microsoft Corporation is a Washington corporation with a principal place of business and headquarters in Redmond, Washington.

FACTUAL ALLEGATIONS

A. Generative AI and Large Language Models

31. The terms "artificial intelligence" or "AI" refer generally to computer systems designed to imitate human cognitive functions.

32. The terms "generative artificial intelligence" or "generative AI" refer specifically to systems that are capable of generating "new" content in response to user inputs called "prompts."

33. For example, the user of a generative AI system capable of generating images from text prompts might input the prompt, "A lawyer working at her desk." The system would then attempt to construct the prompted image. Similarly, the user of a generative AI system capable of generating text from text prompts might input the prompt, "Tell me a story about a lawyer working at her desk." The system would then attempt to generate the prompted text.

34. Recent generative AI systems designed to recognize input text and generate output text are built on "large language models" or "LLMs."

35. LLMs attempt to "understand" human language by processing input text, and are designed to mimic human use of language by generating output text on a predictive basis, i.e., predicting what word follows what.

36. Microsoft's Turing LLMs are a complex web of mathematical functions comprised of a series of algorithms that break down input text into smaller pieces—words or portions of words, called "tokens"—then translate those pieces into "vectors," or a sequence of numbers that is used to identify the token within the series of algorithms. Those vectors help place each token on a map, by identifying other tokens closely associated with the word. "[T]he

process begins by breaking text down into roughly word-length 'tokens,' which are converted to numbers. The model then calculates each token's proximity to other tokens in the training data—essentially, how near one word appears in relation to any other word. These relationships between words reveal which words have similar meanings . . . and functions."[7] As the model trains and digests more expression, the algorithms depicting the relationship between various tokens changes with it.

37. "Training" an LLM refers to the process by which the parameters that define an LLM's behavior are adjusted through the LLM's ingestion and analysis of large "training" datasets.

38. The "training" of an LLM requires inputting large numbers of parameters in the model and then supplying the LLM with large amounts of text for the LLM to ingest—the more text, the better.

39. The model takes text inputs in the form of an incomplete phrase or passage, and attempts to complete the phrase, essentially a fill-in-the-blank quiz. The model compares its predicted phrase completion with the actual "correct" answer. The model then adjusts its internal algorithms to "learn" from its mistakes. In other words, it adjusts its algorithms to reduce the likelihood of making the same mistake again, thus minimizing the difference between any given text input and the "correct" text output.

40. The model then repeats this same cycle millions, possibly billions, of times across the entire corpus, adjusting its algorithms each time to reflect the text input from the corpus. The pre-training process enables the model to process prompts and generate text output that mimics human language. It does so by exposing the model to a wide range of texts and using algorithms

[7] *See* Comment of OpenAI "Re: Notice of Inquiry and Request for Comment [Docket No. 2023-06]," United States Copyright Office, Oct. 30, 2023, p. 5-6 (available at: https://downloads.regulations.gov/COLC-2023-0006-8906/attachment_1.pdf).

to predict the next word in the text. By repeating this process over and over, the model exhibits fluency in style, syntax, and expression of ideas, largely by digesting and processing the expression contained in the material used for training. In this way, the LLM effectively mines and feeds on the expression contained in the training corpus, adjusting its algorithms such that it can mirror and mimic the ordering of words, style, syntax, and presentation of facts, concepts, and themes.

41. After the pre-training process, the generative model generally undergoes a further post-training process. At this point, the model is capable of completing phrases and predicting the next word or words that come next after a particular text input, but cannot yet respond to questions, let alone with human-like responses.[8] The post-training process is sometimes referred to as "finetuning." This stage typically involves more human supervision, and focuses on making adjustments to the model using comparatively smaller training datasets.

42. As the U.S. Patent and Trademark Office has observed, LLM "training" "almost by definition involve[s] the reproduction of entire works or substantial portions thereof."[9]

43. Microsoft itself noted: "Language models with large numbers of parameters, more data, and more training time acquire a richer, more nuanced understanding of language."[10]

44. "Training" in this context is therefore a technical-sounding euphemism for "copying and ingesting expression."

45. Moreover, in some form and to some degree currently unknowable to the public, LLMs "memorize" or store their "training" data (even if in a "translated" form, such as a unique

[8] Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model, Shaden Smith et al., https://arxiv.org/pdf/2201.11990.

[9] U.S. Patent & Trademark Office, *Public Views on Artificial Intelligence and Intellectual Property Policy* 29 (2020), *available at* https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf (last accessed Jan. 22, 2024).

[10] *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model*, Microsoft Research Blog (Oct. 11, 2021), *published at* https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/ (last accessed June 10, 2025).

statistical profile), such that the data (at least in part) can be accessed, recalled, and reproduced by the LLM at will.[11]

46. The quality of the LLM (that is, its capacity to generate human-seeming responses to prompts) is dependent on the quality of the datasets used to "train" the LLM.

47. Books are the high-quality materials Microsoft wants, needs, and has therefore used to develop generative AI products that produce high-quality results: text that appears to have been written by a human writer.

48. Professionally authored, edited, and published books—such as those authored by Plaintiffs here—are an especially important source of LLM "training" data.

49. As one researcher put it: "[large language] model behavior is not determined by architecture, hyperparameters, or optimizer choices [i.e. technical features set during model training]. It's determined by your dataset, nothing else. Everything else is a means to an end in efficiently deliver[ing] compute to approximating that dataset."[12]

50. Another group of AI researchers (not affiliated with Microsoft) observed that "[b]ooks are a rich source of both fine-grained information, how a character, an object or a scene looks like, as well as high-level semantics, what someone is thinking, feeling and how these states evolve through a story."[13]

51. Microsoft researchers wrote in the 2020 release of their Turing-NLG model that "the bigger the model and the more diverse and comprehensive the pretraining data, the better it

[11] *See* Jason Koebler, *Google Researchers' Attack Prompts ChatGPT to Reveal Its Training Data*, 404 Media (Nov. 29, 2023), *available at* https://www.404media.co/google-researchers-attack-convinces-chatgpt-to-reveal-its-training-data/ (last accessed Jan. 22, 2024); Kent K. Chang *et al.*, *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4* (2023), *available at* https://arxiv.org/pdf/2305.00118v1.pdf (last accessed Jan. 22, 2024).
[12] *See* "The 'it' in AI models is the data set," James Betker, https://nonint.com/2023/06/10/the-itin-ai-models-is-the-dataset/ (June 10, 2023) (emphasis added).
[13] Yukun Zhu *et al.*, *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books* 1 (2015), *available at* https://arxiv.org/pdf/1506.06724.pdf (last accessed Jan. 22, 2024).

performs at generalizing to multiple downstream tasks even with fewer training examples."[14] And, in the 2022 paper announcing Microsoft's follow-up Megatron-Turing NLG model, Microsoft and Nvidia researchers similarly filtered their webscraped documents for "high-quality" by training a classifier to label documents most similar to "OpenWebText2, Wikipedia, and Books3," as well as using Books3 itself as part of the training dataset.[15]

52. Once the "training" data is ingested, the creator of an LLM can then control how closely the LLMs' outputs adhere to probability. Software engineers refer to this parameter as "temperature." If engineers set an LLMs at a "hotter" temperature, the model will bias *against* what it calculates as the most probable response in favor of more random outputs. Likewise, the "cooler" the LLMs are set, the more closely their outputs will adhere to statistical probability. In this way, the creator of an LLM can control the very perception of copying in the outputs of an LLM.

53. Once an LLM is trained on Plaintiffs' and the proposed class's works, it cannot function without exploiting the expression extracted from Plaintiffs' works and retained inside the models. As such, LLMs are themselves infringing copies, made without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act.

B. Microsoft Engaged in Largescale Copyright Theft in Training Its Megatron LLM

1. Microsoft's "Turing" LLMs and Megatron

54. Microsoft developed several LLMs, including the Megatron LLM, as part of its "Project Turing."

[14] Turing-NLG: A 17-billion-parameter language model by Microsoft," Rosset, Cory, Microsoft Research, Feb. 13, 2020, https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/.

[15] Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model, Smith et al., https://arxiv.org/pdf/2201.11990. The authors noted that "[c]areful processing of high-quality, high-volume and diverse datasets directly contributes to model performance in downstream tasks." *Id.* at 2.

55. Project Turing is a "deep learning initiative inside Microsoft to build the best-in-class models for use by Microsoft and power AI applications across the entire Microsoft product family (Word, PowerPoint, Office, Dynamics, etc.) and make them available for use through Azure."[16]

56. Microsoft announced "Project Turing" in 2017 and stated that its goal was to "evolve Microsoft products with the adoption of deep learning for both text and image processing."[17] Its work was to be "integrated into multiple Microsoft products including Bing, Office, and Xbox."[18]

57. In February 2020, Project Turing released its first LLM, the "Turing-NLG" LLM—at the time the "largest model ever published at 17 billion parameters."[19] In the post announcing Turning-NLG, Microsoft previewed that the model would soon power "experiences with the Microsoft Office suite by offering writing assistance to authors and answering questions that readers may ask about a document."[20]

58. The "Turing-NLG" model was surpassed by an LLM created by OpenAI, GPT-3.

59. In October 2021, Microsoft released Megatron, which was trained to some 530 billion parameters, a 31-fold increase over Turing-NLG.[21] Megatron set the record for largest language model upon its release. And, as with its Turing-NLG predecessor, Microsoft noted that

---

[16] *See* https://www.microsoft.com/en-us/research/project/project-turing/ (last accessed June 10, 2025).
[17] *Turing-NLG: A 17-billion-parameter language model by Microsoft*, Microsoft Research Blog (Feb. 13, 2020), *published at* https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/ (last accessed Mar. 17, 2025); "Inside Microsoft's Project Turing, the team that's quietly reinventing how it develops advanced AI to move faster and take on rivals like Google," *Business Insider,* Oct. 12, 2021, https://www.businessinsider.com/microsoft-project-turing-ai-large-language-models-google-openai-2021-9 (last accessed Mar. 17, 2025).
[18] *Id.*
[19] *Turing-NLG: A 17-billion-parameter language model by Microsoft*, Microsoft Research Blog (Feb. 13, 2020), *published at* https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/ (last accessed Mar. 17, 2025).
[20] *Id.*
[21] *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model*, Microsoft Research Blog (Oct. 11, 2021), *published at* https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/ (last accessed June 10, 2025).

the innovations in Megatron were set to benefit "tomorrow's products."[22] Unsaid, however, was that these benefits were secured from the unauthorized exploitation of Plaintiffs' works.

60. Microsoft will reap massive commercial advantage from incorporating these models into "tomorrow's products." Already, Microsoft has boasted of the added features and charges, surges in usage, and new revenue it has reaped from incorporating OpenAI's LLMs into its products. Analysts project that just one integration of LLMs into Microsoft's products could generate more than $10 billion in annualized revenue by 2026. Microsoft will see similar commercial benefits from the LLMs it develops itself.

2. Microsoft Copied a Massive Trove of Pirated Books To Train Megatron

61. Microsoft has publicly acknowledged that it used subsets of a dataset called the Pile. The Pile is an 800 GB+ open-source dataset created for large language model training. The Pile was hosted and made publicly available online by a nonprofit called EleutherAI. As described by its creators, "The Pile is constructed from 22 diverse high-quality subsets . . . many of which derive from academic and professional sources. . . . [M]odels trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, while improving performance on downstream evaluations."[23]

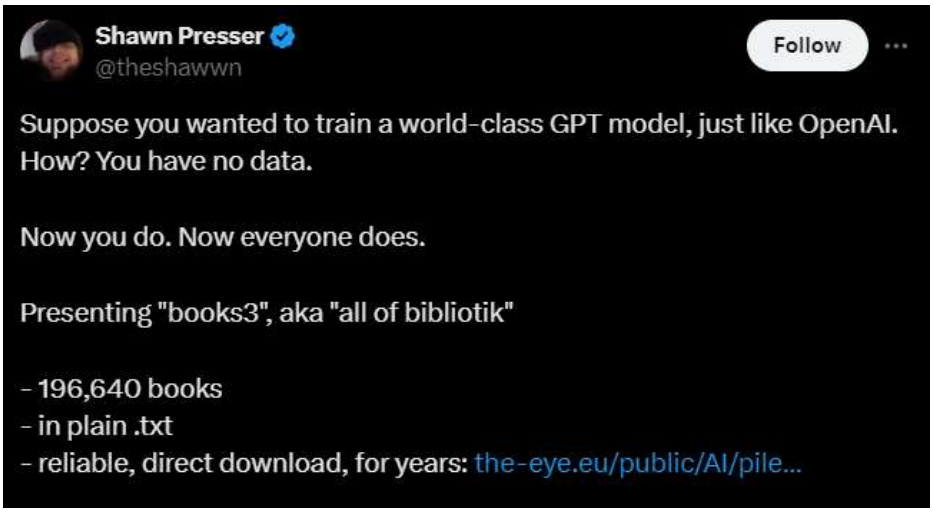
62. The Pile includes the notorious pirated dataset Books3, comprising nearly 200,000 books. Books3 is one of the Pile datasets Microsoft used to train Megatron.[24] To put a finer point on it: Microsoft trained Megatron on 25.7 billion tokens of Books3.[25]

---

[22] *Id.*
[23] EleutherAI, "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," (Dec. 31, 2020) https://arxiv.org/pdf/2101.00027.
[24] *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model*, Microsoft Research Blog (Oct. 11, 2021), *published at* https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/ (last accessed June 10, 2025).
[25] *Id.*

63. Microsoft selected this subset from the Pile because they "found [it] to be of the highest relative quality."[26]

64. Books3 was created by an independent developer named Shawn Presser who is also one of The Pile's architects. Presser described how he created Books3 in a Twitter thread from October 2020:[27]



65. Presser went on. He said he created Books3 in response to "OpenAI's papers on GPT-2 and 3," two of OpenAI's LLMs. Those papers, Presser noted, "refer[] to datasets named 'books1' and 'books2,'" the latter of which Presser suspects "might be 'all of libgen.'"[28] LibGen refers to "Library Genesis," a website offering pirated books that was ordered shut down for copyright infringement in 2015. *See Elsevier, Inc. et al v. www.Sci-hub.org et al*, 15-cv-2482-RWS, Dkt. 53 (Oct. 30, 2015). In other words, OpenAI at that time was training on datasets called Books1 and Books2, which Presser suspected might be created from pirated books datasets. As later transpired, Presser was correct.

[26] *Id*.
[27] *See* Tweet by Shawn Presser, Oct. 25, 2020, https://x.com/theshawwn/status/1320282149329784833?lang=en (last accessed June 10, 2025).
[28] *See* Tweet by Shawn Presser, Oct. 25, 2020, https://x.com/theshawwn/status/1320282152689336320 (last accessed June 10, 2025).

66. At that time, moreover, Microsoft was already an investor in and partner of OpenAI. In other words, Presser's creation of a vast trove of pirated books was a reaction to the piracy that Microsoft itself helped enable.

67. Presser sought to create a pirated-book dataset comparable to what he suspected OpenAI created for itself. Presser announced that Books3 was also a direct download of all books from a different pirated website—a compilation of "**196,640 books**," which comprises "**all of bibliotik**."[29]

68. Bibliotik is a "notorious pirated collection" of "pirated books."[30] For years prior to its use as "Books3," Bibliotik was frequently included in roundups of the best—and most popular—sources for pirated material.[31]

69. Books3 was a critical part of The Pile. In EleutherAI's paper on The Pile, it touted the key value of Books3 as training material: "Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker . . . Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude larger than our next largest book dataset

---

[29] *See* Tweet by Shawn Presser, Oct. 25, 2020, https://x.com/theshawwn/status/1320282149329784833?lang=en (last accessed June 10, 2025).

[30] *See* Schoppert, "Whether you're an undergraduate doing research, or a fan of the Nick Stone novels, or indeed a hungry AI...," Nov. 29, 2022, https://aicopyright.substack.com/p/whetheryoure-an-undergraduate-doing ("What is Bibliotik? A notorious pirated collection."); "What I Found in a Database Meta Uses to Train Generative AI," Alex Reisner, *The Atlantic*, Sept. 25, 2023, https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-metacopyright-infringement-lawsuit/675411/ ("a collection of pirated ebooks, most of them published in the past 20 years."); "Revealed: The Authors Whose Pirated Books are Powering Generative AI," Alex Reisner, *The Atlantic*, Aug. 19, 2023, https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-piratedbooks/675063/ ("collections of pirated books, such as Library Genesis, Z-Library, and Bibliotik, that circulate via the BitTorrent file-sharing network."); "Are ChatGPT, Bard and Dolly 2.0 Trained On Pirated Content?," Roger Monti, *Search Engine Journal*, April 20, 2023, https://www.searchenginejournal.com/are-chatgpt-bard-and-dolly-2-0-trained-on-piratedcontent/485089/ ("The Books3 dataset contains the text of books that were pirated and hosted at a pirate site called, bibliotik.").

[31] *See* Commit History of "Awesome Piracy," Github.com, Oct. 13, 2018, https://github.com/aviranzerioniac/awesomepiracy/commit/61928765e3ee0b4f3dbe3c0724b196e5f0f17e59?short_path=5a831ea#diff-5a831ea67cf5cf8703b0de46901ab25bd191f56b320053be9332d9a3b0d01d15 (October 13, 2018 commit to "awesome piracy" repo listing "Bibliotik Popular ebooks/audiobooks private tracker"); "Reddit Piracy Megathread" repo., Github.com, Mar. 21, 2019, https://github.com/magicoflolis/Reddit-Piracy-Megathread/blob/master/data/findingtextbooks.md (March 21, 2019 guide from "r/piracy" on how to source textbooks listing "Bibliotik"); "List of free eBook download sites," Pirates-forum.org, Mar. 6, 2014, https://pirates-forum.org/Thread-List-of-free-eBook-downloadsites?highlight=bibliotik (March 31, 2014 post from "pirates forum" thread entitled "List of free eBook download sites" listing "bibliotik").

(BookCorpus2)." The paper then summarized its key point for why The Pile included this known source of illegal copyright material: "**We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling**."[32] (emphasis added).

70. At the same time, Presser and EleutherAI repeatedly and publicly acknowledged that, with The Pile and Books3, they were making available a cache of pirated material. EleutherAI's paper on The Pile noted that "there is little acknowledgment of the fact that the processing and distribution of data owned by others may also be a violation of copyright law."[33] Furthermore, The Pile's datasheet notes that "Books3 is almost entirely comprised of copyrighted works . . . ."[34] Presser, for his part, has admitted to releasing Books3 despite "fear of copyright backlash."[35]

71. In August 2023, Books3 was removed from the "most official" copy of The Pile hosted by "The Eye" due to copyright complaints. Despite this takedown, the original version appears otherwise available as part of The Pile from other sources.

72. Microsoft, in training Megatron on Books3 and other subsets of The Pile, has taken authors' works contained in these datasets without compensation, and has deprived authors of books sales and licensing revenues. There has long been an established market for the sale of books and e-books. Yet Microsoft ignored that market and chose to download a massive corpus of copyrighted books from the internet, without even paying for an initial copy. Microsoft has not revealed whether it has downloaded other pirated book datasets, which also would infringe to the extent that they made separate downloads apart from Books3.

---

[32] "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," Gao et al, p. 3-4, https://arxiv.org/pdf/2101.00027 (last accessed June 10, 2025).
[33] *Id*. at 14-15.
[34] "Datasheet for the Pile," Gao et al, Jan. 20, 2022, p. 15, https://arxiv.org/pdf/2201.07311 (last accessed June 10, 2025).
[35] Comment of "sillysaurusx," Hacker News, Jul. 11, 2023, https://news.ycombinator.com/item?id=36685115 (last accessed June 10, 2025).

73. Microsoft has also usurped a licensing market for copyright owners. In the last two years, a thriving licensing market for copyrighted training data has developed. A number of AI companies, including Microsoft as well as OpenAI, Google, and Meta, have paid hundreds of millions of dollars to obtain licenses to reproduce copyrighted material for LLM training. These include deals with HarperCollins (Microsoft), Axel Springer, News Corporation, the Associated Press, and others. Furthermore, absent Microsoft's largescale copyright infringement, blanket licensing practices would be possible through clearinghouses, like the Copyright Clearance Center, which recently launched a collective licensing mechanism that is available on the market today.[36]

74. Microsoft, however, has chosen to use Plaintiffs works and the works owned by the Class free of charge, and in doing so has harmed the market for the copyrighted works and depriving them of book sales and licensing revenue.

C. Harm to Authors

75. LLMs like Megatron seriously threatens the livelihood of the very authors—including Plaintiffs here, as discussed specifically below—on whose works Megatron was "trained" without the authors' consent.

76. Goldman Sachs estimates that generative AI could replace 300 million full-time jobs in the near future, or one-fourth of the labor currently performed in the United States and Europe. *See also* "Behind the Curtain: A White-Collar Bloodbath," *Axios,* May 28, 2025, https://www.axios.com/2025/05/28/ai-jobs-white-collar-unemployment-anthropic ("Dario Amodei — CEO of Anthropic, one of the world's most powerful creators of artificial intelligence — has a blunt, scary warning for the U.S. government and all of us: AI could wipe

[36] Copyright Clearance Center, The Intersection of AI & Copyright, https://www.copyright.com/resource-library/insights/intersection-ai-copyright/ (last visited June 10, 2025).

out half of all entry-level white-collar jobs — and spike unemployment to 10-20% in the next one to five years . . . .").

77. Already, writers report losing income from copywriting, journalism, and online content writing—important sources of income for many book authors. The Authors Guild's, the oldest professional organization representing writers and authors, published an earnings report for 2023 showing a median writing-related income for full-time authors of just over $20,000, and that full-time traditional authors earn only half of that from their books.[37] The rest comes from activities like content writing—work that is starting to dry up as a result of generative AI systems trained on those writers' works, without compensation, to begin with.

78. Other examples abound. *See, e.g.*, Jin Liu, Xingchen Xu, Xi Nan, Yongjun Li, and Yong Tan, "Generate" the Future of Work through AI: Empirical Evidence from Online Labor Markets, https://arxiv.org/abs/2308.05201 (last revised June 18, 2025); Mauran, Cecily, "If you're buying the Kara Swisher book on Amazon, make sure it's not AI-generated knockoff," *Mashable,* February 28, 2024, https://mashable.com/article/kara-swisher-book-amazon-ai-generated-knockoff.; Oremus, Will, "Tech writer Kara Swisher has a new book. Enter the AI-generated scams." *The Washington Post*, March 1, 2024, https://www.washingtonpost.com/technology/2024/03/01/amazon-ai-fake-books-authors/; Roscoe, Jules, "AI-Generated Books of Nonsense Are All Over Amazon's Bestseller Lists," *Vice,* June 28, 2023, https://www.vice.com/en/article/ai-generated-books-of-nonsense-are-all-over-amazons-bestseller-lists/.

79. Microsoft's unauthorized commercial copying of Plaintiffs' works and works owned by the proposed Classes was manifestly unfair use. Microsoft copied Plaintiffs' books

[37] Authors Guild, "Top Takeaways from the 2023 Author Income Survey (2023), *available at* https://authorsguild.org/news/key-takeaways-from-2023-author-income-survey/#:~:text=Though%20overall%20author%20incomes%20are,coming%20in%20a%20close%20second (last accessed Jan. 22, 2024).

without compensation. It has usurped authors' content for the purpose of creating a machine built to generate the very type of content for which authors usually would be paid.

80. In short, the success of Microsoft's Megatron LLM is predicated on mass copyright infringement.

D. Microsoft Exploited Each of Plaintiffs' Copyrighted Works

81. The continuing commercial viability of Plaintiffs' works is endangered by Microsoft.

82. Each author represented here has a distinct voice, a distinct style, and distinct creative expression. But all Plaintiffs have suffered identical harms from Microsoft's infringing reproductions of their works.

83. Plaintiffs make the specific allegations of infringement below based on what is known about Microsoft's training practices; what is known about the contents, uses, and availability of the pirate book repositories such as The Pile and Books3.

84. Many of the works authored and owned by Plaintiffs, including Alter, Bird, Bly, LaValle, Linden, Okrent, Sides, Tolentino, Vail, and Winchester, are available on Books3.

85. ***Plaintiff Alter.*** Alter is the author of a number of New York Times bestsellers, including The Center Holds: Obama and His Enemies; The Promise: President Obama, Year One; and The Defining Moment: FDR's Hundred Days and the Triumph of Hope. Each of those three books are a part of the Books3 dataset. Pirated copies of each of those three books—as well as most recent book *His Very Best: Jimmy Carter, A Life*—are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik. Alter is the author and owner of the registered copyrights listed under his name in Exhibit A.

86. ***Plaintiff Bird.*** Bird is the recipient of the 2006 Pulitzer Prize for Biography for *American Prometheus: The Triumph and Tragedy of J. Robert Oppenheimer*. A number of this books, including *American Prometheus*, *The Good Spy*, and *The Color of Truth* are a part of the

Books3 dataset. Pirated copies of each of all of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik. Bird is the author and owner of the registered copyrights listed under his name in Exhibit A.

87. ***Plaintiff Linden.*** Linden is the author of nine nonfiction books, including *The Parrot's Lament*, and *Other True Tales of Animal Intrigue, Intelligence, and Ingenuity; The Octopus and the Orangutan: More True Tales of Animal Intrigue, Intelligence, and Ingenuity; The Alms Race: The Impact of American Voluntary Aid Abroad*; and *Winds of Change*. Two of his books, *The Parrot's Lament* and *The Ragged Edge of the World*, are a part of the Books3 dataset. Pirated copies of a number of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik. Linden is the author and owner of the registered copyrights listed under his name in Exhibit A.

88. ***Plaintiff Okrent***. Okrent is the author of a number of nonfiction books, including *Great Fortune: The Epic of Rockefeller Center*, *Last Call: The Rise and Fall of Prohibition*, and *The Guarded Gate: Bigotry, Eugenics and the Law that Kept Two Generations of Jews, Italians and Other European Immigrants Out of America.*

89. At least three of his books, including *Last Call* and *The Guarded Gate*, are a part of the Books3 dataset. Pirated copies of a number of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik. Okrent is the author and owner of the registered copyrights listed under his name in Exhibit A.

90. ***Plaintiff Sides***. Sides is the author of a number of *New York Times* bestsellers, including *Ghost Soldiers: The Epic Account of World War II's Greatest Rescue Mission* and *Blood and Thunder: An Epic of the American West*. He is an editor-at-large for *Outside* and is a frequent contributor to *National Geographic*.

91. Six of his books, including *Ghost Soldiers* and *Blood and Thunder*, are a part of the Books3 dataset. Pirated copies of all of his books are available on the internet through

websites like LibGen, ZLibrary, and/or Bibliotik. Sides is the author and owner of the registered copyrights listed under his name in Exhibit A.

92. ***Plaintiff Tolentino***. She is the author of the *New York Times* bestseller *Trick Mirror: Reflections on Self-Delusion*. She is a staff writer for *The New Yorker* whose work has also appeared in *The New York Times Magazine* and *Pitchfork.*

93. Her book *Trick Mirror is* part of the Books3 dataset. Pirated copies of it are widely available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik. Tolentino is the author and owner of the registered copyright listed under her name in Exhibit A.

94. ***Plaintiff Winchester***. Winchester is the author of several *New York Times* bestsellers, including *The Professor and the Madman*; *The Map That Changed the World: William Smith and the Birth of Modern Geology*; and *The Men Who United the States: America's Explorers, Inventors, Eccentrics, and Mavericks, and the Creation of One Nation, Indivisible*.

95. At least eleven of his books, including *The Professor and the Madman* and *Krakatoa*, are a part of the Books3 dataset. Pirated copies of nearly all of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik. Winchester is the author and owner of the registered copyrights listed under his name in Exhibit A.

96. ***Plaintiffs Bly and Eloisa James, Inc.*** Bly is a tenured professor and former chair of the English department at Fordham University who also writes best-selling Regency and Georgian romance novels under the pen name Eloisa James. Some of Bly's most popular works include books in the *Desperate Duchesses* series, the *Fairy Tales* series, the *Wildes of Lindow Castle* series, and the *Essex* series. Eloisa James, Inc. is a loan-out corporation owned by Bly.

97. Bly is the author of and owner of the registered copyrights under her name in Exhibit A. Bly is also the author of the registered copyrights listed under Eloisa James, Inc. in Exhibit A. Eloisa James, Inc. is the legal and/or beneficial owner of the registered copyrights listed under its name in Exhibit A.

98. Upon information and belief, Bly's works *An Affair Before Christmas* and *A Gentleman Never Tells* (and others) appear in the pirated Books3 dataset.

99. Microsoft ingested and copied all or many of the works listed under Bly and Eloisa James, Inc. in Exhibit A without permission (the "Bly Infringed Works").

100. ***Plaintiff LaValle***. LaValle is an associate professor of Creative Writing at Columbia University and the author of five novels, a short story collection, two novellas, and two comic books. Some of LaValle's most popular novels include *Big Machine, The Devil in Silver*, and *The Changeling*.

101. LaValle is the sole author of and owner or beneficial owner of the registered copyrights in six (6) written works of fiction contained in Exhibit A to this Complaint, at 6.

102. Upon information and belief, LaValle's works *The Devil in Silver* and *The Changeling* (and others) appear in the pirated Books3 dataset.

103. Microsoft ingested and copied all or many of the works listed under LaValle in Exhibit A without permission (the "LaValle Infringed Works").

104. ***Plaintiff Vail***. Rachel Vail is an award-winning American author who primarily authors children's and young adult books. Some of Vail's most popular novels include *Ever After*, *Unfriended*, and *Justin Case: School, Drool, and Other Daily Disasters*.

105. Vail is the sole author of and owner or beneficial owner of the registered copyrights in twenty-four (24) written works of fiction contained in Exhibit A to this Complaint, at 8–9.

106. Upon information and belief, Vail's work *Unfriended* appears in the pirated Books3 dataset.

107. Microsoft ingested and copied all or many of the works listed under Vail in Exhibit A without permission (the "Vail Infringed Works").

CLASS ALLEGATIONS

A. Class Definitions

108. This action is brought by Plaintiffs individually and on behalf of the following Class, as defined below, pursuant to Rule 23(b)(3) and 23(b)(1) of the Federal Rules of Civil Procedure.

109. The Class consists of:

> All legal or beneficial owners of copyrighted works that: (A) were registered with the United States Copyright Office within five years of the work's first publication; (B) were downloaded or otherwise reproduced by Microsoft; (C) were registered with the United States Copyright Office before being downloaded or otherwise reproduced Microsoft, or were registered within three months of first publication; and (D) are assigned one or more International Standard Books Number(s) (ISBN) or Amazon Standard Identification Number(s) (ASIN).

110. The Class excludes (1) the judge assigned to these proceedings, their staff and their immediate family members; and (2) Defendant, its officers and directors, members of their immediate families, their co-conspirators, aiders and abettors, and the heirs, successors or assigns of any of the foregoing.

B. Rules 23(a) and 23(g)

111. The Class consists of at least tens of thousands of authors and copyright holders and thus are so numerous that joinder of all members is impractical.

112. The identities of members of the Class can be readily ascertained from business records maintained by Microsoft and at a minimum from the content of the Books3 database that Microsoft illegally downloaded.

113. The claims asserted by Plaintiffs are typical of the claims of the Class because their copyrights were infringed in materially the same way and their interests in preventing future infringement and redressing past infringement are materially the same.

114. The Plaintiffs will fairly and adequately protect the interests of the Class and do not have any interests antagonistic to those of other members of the Class.

115. Plaintiffs have retained attorneys who are knowledgeable and experienced in copyright and class action matters, as well as complex litigation.

116. There are questions of fact or law common to the Class, including:

a. Whether Microsoft's copied works owned by Plaintiffs and the members of the Class;

b. Whether Microsoft's copying of Plaintiffs' and the Class's copyrighted works consisted direct infringement; and

c. Whether Microsoft's copying of works owned by Plaintiffs and the Class was willful.

C. Rule 23(b)

117. Microsoft have acted on grounds common to Plaintiffs and the Class by treating all Plaintiffs' and Class Members' works equally, in all material respects, in their LLM "training."

118. Common questions of liability for infringement predominate over any individualized damages determinations as may be necessary. To decide liability, the Court will necessarily apply the same law to the same conduct, which Microsoft engaged in indiscriminately with respect to all Plaintiffs and all members of the Class.

119. Further, to the extent Plaintiffs elect to pursue statutory rather than actual damages before final judgment, the damages inquiry will likewise be common, if not identical, across Plaintiffs and members of the Class.

120. A class action is superior to any individual litigation of Plaintiffs' and Class Members' claims. Class Members have little interest, distinct from Plaintiffs' and other Class Members', in prosecuting individual actions. It would waste judicial resources to decide the same legal questions repeatedly, thousands of times over, on materially indistinguishable facts. The Classes presents no special manageability problems.

121. This action is also appropriate as a class action pursuant to Rule 23(b)(2) of the Federal Rules of Civil Procedure because Microsoft infringing conduct is applicable generally to Plaintiffs and the proposed Class and the requested injunctive relief is appropriate respecting the proposed Class as a whole.

D. Rule 23(c)(4)

122. In the alternative to certification under Rule 23(b)(3) and 23(b)(2), common questions predominate within the determination of liability for infringement, therefore the issue of liability may be separately certified for class treatment even if the entire action is not.

CLAIM FOR RELIEF

COPYRIGHT INFRINGEMENT (17 U.S.C. § 501)

123. Plaintiffs and members of the Class incorporate by reference the allegations in Paragraphs 1 to 122 as though fully set forth herein.

124. Plaintiffs and members of the Class own the registered copyrights in the works that Microsoft reproduced and appropriated to train their artificial intelligence models.

125. Plaintiffs and members of the Class therefore hold the exclusive rights, including the rights of reproduction and distribution, to those works under 17 U.S.C. § 106.

126. Microsoft infringed on the exclusive rights, under 17 U.S.C. § 106, of Plaintiffs and members of the proposed Class by, among other things, reproducing the works owned by Plaintiffs and the proposed Class, namely, by downloading from pirated sources, and, separately, by training on reproduced copies of works.

127. On information and belief, Microsoft's infringing conduct alleged herein was and continues to be willful. Microsoft infringed on the exclusive rights of Plaintiffs and members of the proposed Class knowing that its conduct was infringing.

128. Plaintiffs and members of the proposed Class are entitled to statutory damages, actual damages, disgorgement, and other remedies available under the Copyright Act.

129. Plaintiffs and members of the proposed Class have been and continue to be irreparably injured due to Microsoft's conduct, for which there is no adequate remedy at law. Microsoft will continue to infringe on the exclusive right of Plaintiffs and the proposed class unless their infringing activity is enjoined by this Court. Plaintiffs are therefore entitled to permanent injunctive relief barring Microsoft's ongoing infringement.

PRAYER FOR RELIEF

130. Plaintiffs, on behalf of themselves and all others similarly situated, pray for the following relief:

a. Certification of this action as a class action under Federal Rule of Civil Procedure 23;

b. Designation of Plaintiffs as class representatives;

c. Designation of Plaintiffs' counsel as class counsel;

d. An injunction prohibiting Microsoft from infringing on Plaintiffs' and class members' copyrights, including without limitation enjoining Microsoft from taking copyright protected books from pirated sources and using Plaintiffs' and class members' copyrighted works in "training" Microsoft's large language model without express authorization;

e. An award of actual damages to Plaintiffs and class members;

f. An award of Microsoft's additional profits attributable to infringement to Plaintiffs and class members;

g. An award of statutory damages up to $150,000 per infringed work to Plaintiffs and members of the Class, in the alternative to actual damages and profits, at Plaintiffs' election before final judgment;

h. Reasonable attorneys' fees and costs, as allowed by law;

i. Pre-judgment and post-judgment interest, as allowed by law; and

j. Such further relief as the Court may deem just and proper.

DEMAND FOR JURY TRIAL

Pursuant to Rule 38 of the Federal Rules of Civil Procedure, Plaintiffs hereby demand a jury trial for all claims so triable.

Dated: June 24, 2025

/s/ Rachel Geman
Rachel Geman
Anna J. Freymann
Wesley Dozier*
Danna Elmasry
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, NY 10013
Tel.: 212.355.9500
rgeman@lchb.com
afreymann@lchb.com
wdozier@lchb.com
delmasry@lchb.com

/s/ Rohit D. Nath
Justin A. Nelson*
Alejandra C. Salinas*
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Tel.: 713.651.9366
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath*
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067
Tel.: 310.789.3100
rnath@susmangodfrey.com

/s/ Scott J. Shoulder
Scott J. Shoulder
CeCe M. Cole
COWAN DEBAETS ABRAHAMS & SHEPPARD LLP
60 Broad Street, 30th Floor
New York, NY 10010
Tel.: 212.974.7474
ssholder@cdas.com
ccole@cdas@com

J. Craig Smyser
Charlotte Lepic
SUSMAN GODFREY L.L.P.
One Manhattan West, 51st Floor
New York, NY 10001
Tel.: 212.336.8330
csmyser@susmangodfrey.com
clepic@susmangodfrey.com

Counsel for Plaintiffs and the Proposed Class

**Pro hac vice forthcoming*